ESCAPE WP4 Meeting - zoom on line meeting- 2020 Sept 7-8

Presentation of projects use-cases

9. VizieR catalogs
Gilles Landais
07/09/2020, 16:00-16:20

origin needed to get CoreTrustSeal
"on-top" provenance from available metadata
traceability

- entities: Vizier catalogs, articles, tables (original, transformed)
- descriptions of entities: VOTable, ReadMe
- agents: CDS, A&A..., authors
- activities: copy, transform, enrich with metadata
- identifiers: DOI, ivoID

Serializations:
* VODMLLite annotation within VOTable , (renamed as ModelInstanceinVotable)
* yaml
--> "intermediate" serialization that is readable for the user and compatible with the data model (i.e. machine readable), could be transformed to graphs and W3C
--> put in the VOTable, or link to it
http://cdsarc.u-strasbg.fr/viz-bin/provenance?cat=I/345&out=vodmlite&filter=true
ML/ after meeting notes :
Curation workflow: (ingestion in Vizier)
suggestion : bind together in a collection entity : Readme , original data from A&A , published Article
VOmetadata decoration workflow: (added-value metadata)
filter metadata annotation, link to SVO filter profile service, conversion to other photometric system

12. Multi-frequency polarimetry of a complete sample of extragalactic radio sources
Vincenzo Galluzzi (INAF - Osservatorio Astronomico di Trieste)
07/09/2020, 16:20-16:40

Complex process, workflow, large number of calibration files
"on-top" mainly, having a worklflow (several possible paths though),
reproducibility

- entities: ATCA and ALMA data products (Level 0-1-2-3-4), ancillary products
- descriptions of entities: ASDM, or formats defined by projects (ALMA, ATCA), png.jpeg, logs/text
- activities : level 0-->1, 1-->2...
- description of activities: code on gitlab, IDL, bash scripts (for MIRIAD) and Python scripts (for CASA)

granularity needed for the end user, e.g. for data discovery a tree diagram for data products with some quality metrics metadata, while for debugging/reproducing the granularity should be at the level of each calibration or imaging task (level 1-->2 and 2-->3)
Many steps involved, need to centralize information: in order to ease such a process, we are considering the porting to python (python scripts for all calibration/imaging, jupyter notebooks for generating level 4 data products)

15. KM3NeT (TBC)
Jutta Schnabel (Friedrich-Alexander Universität Erlangen-Nürnberg), Tamas Gal (ECAP / FAU)
07/09/2020, 16:40-17:00

Inside provenance

- activities: data processing steps, event simulation
    - km3pipe: include provenance dictionary with inputs/outputs (json)
- description of activities: CWL, DIRAC

    - --> is it sufficient to expose prov?
- entities: data levels well defined (1->4), primary product = reconstructed neutrino event samples
- descriptions of entities: hdf5, ROOT, ascii, VOEvent

DB provenance (at acquisition): config, calib, trigger params, other parameters
in-file provenance (UUID, environment, time)

after session comments:
MireilleL: it seems the KM3net provenance is proposed for 2 kinds of actors :
- scientists who will look for the scientific parameters applied for activities and the datasets used and generated (ex: nb of iterations....), in addition to activity descriptions and results datasets
- data engineers who will check the quality of execution, the validity of version in an activity description, the execution duration, execution environment, etc. , the size, presence absence of some datasets.
Are some provenance metadata valid for both kinds? What is the overlap?
What is only relevant for science interpretation?
only for workflow execution?

13. LST: Large Size Telescope prototype for CTA
Jose Enrique Ruiz
08/09/2020, 09:30

LST1 in operation, 2 campaigns, detection of the Crab with pulses
Case of capture of provenance : post processing of log files, captured with logprov
* descriptions of activities: processing with: ctapipe (framework) + lstchain (library) + LSTOSA (pipeline)
* entities: DL0-->DL1-->DL2
pipeline run on a grid, parallel jobs
several logprov files collected, with copy of config files
post-processing to extract high level provenance
Storing provenance in a NOSQL database could be a solution of storing

ML: Post meeting :
Observations are composed products : for instance 1 obs --> 10 runs each with 3 to 10 sub-runs (?) of 40 mns
Provenance metadata is filtered from log files, and adjusted to the appropriate granularity to produce JSON serialization documents.
JSON docs are easy to ingest then into MongoDB , for instance .
Storage of parameters is done with the data .
Question : how interpret the provenance records ? to browse the graphs ? need an interface to analyse it.

14. BASS2000: Solar Survey Archive
Jean Aboudarham (LESIA/PADC - Observatoire de Paris/PSL)
08/09/2020, 09:50

3 types of data : observations (images), features (catalogs?), spectra
Observations:
- information in the name of the file and in the header of the fits (instrument name)
Features:

- Provenance requirement: data origin and code processing
- Already provided: instrument, used code (name and software)

Spectra:
- visible atlas spectra: difficult to know provenance information (no agent for example)
ML: a kind of provenance to build afterwards , from the existing info- (a posteriori) .
but you can still fill in the activities you can identify, in this case: at least observing activity , processing activity.

JER: in LST there is a tree-folder structure to save datasets that embeds some provenance params : folder with a date of observations , scripts versions, etc ..
some encoding using names may exist in your project .

17. HEK database: solar events using VOEvent
Veronique Delouille (STCE/Royal Observatory of Belgium)
08/09/2020, 10:10

Detection of solar features with 2 modules run on a complex architecture
* entities: data from SDO satellite, pixel-wise classification maps, region maps, VOevents (output)
* descriptions of activities: SPoCA-AR, SPoCA-CH
* activities: pixel-wise classification, building of maps, tracking from 1 map to the other
* + config
requirement: high level provenance
recurrent request for "more info about intermediary products" (classification maps, region maps)
looking for inclusion of Provenance information in the VOEvent document. --> "How" section in VOEvent
What are the benefits of provenance implementation versus time to organize it? --> start with general, "on-top" provenance, then describe more activities, then try the "inside"approach (but time consuming)
This use case looks like the VizieR use case: add-on of Prov in the dataset.

16. LOFAR / APERTIF / WSRT surveys (TBC)
Mr Yan Grange (ASTRON, the Netherlands Institute for Radio Astronomy)
ASTRON Data from 3 interferometers: WSRT, LOFAR, APERTIF
Long-term archives (LTA) with metadata

* activities: example flow for calibrators and targets, flagged, then calibrated

Provenance:
* inside data files
* ALTA (Apertif) : archive based on IVOA ProvDM

User access: provenance button which allows user to see the intermediate entities

10. European VLBI Network (EVN) Archive
Des Small (JIVE), Harro Verkouter (Joint Institute for VLBI ERIC), Mark Kettenis (JIVE)
08/09/2020, 11:10

* entities: FITS-IDI files, ascii tables (amplitude calib, flagging table), logs, images/plots (output)

* level 0: .cor + logs
* level 1: FITS-IDI, tables
* level 2-3: images/plots

* pipeline: 1-->2-3 (EVN.py)
stable archive (no need to rerun pipeline in 20 years)
"pipeline" outside of control: not automatic, human choices (different strategies)
now will use Jupyter Notebooks --> 1 activity (fixes the granularity), but need to locate the inputs/outputs/parameters
ML: running a Jupyter note book = activity
content of Jupyter notebook = ActivityDescription
is there a need for versioning the note books ?

JER: Is there a need to capture "provenance/history of the activity/notebooks" instead/as well as of "provenance of data"? Could this be solved with kind-of git diffing on notebooks?

18. NenuFAR / ExPRES
Alan Loh, Mr Baptiste Cecconi (Observatoire de Paris)
08/09/2020, 11:30
ExPRES code:
IDL code, planetary radio observations
main goal: reproducibility
entities: magnetic model file, observer location, cdf files (output), PNG files (output)
parameters & config files
activities: expres-dev, expres-quicklook
Provenance managed by OPUS? => how to export Provenance from OPUS process (e.g. IDL script) and insert into global provenance ?
NenuFAR

description of activities: workflows available, different backends and steps (IDL and/or Python)
entities: beamformer mode FITS, images, config/calib files

Question FB : what is the main goal of provenance tracking in Nenufar?
Ans: - quality of results

- tracking all steps and parameters introduced to support reproducibility.

raw data not kept => reproducibility can't be done from first data
Provenance info would be kept after raw data is deleted
JER : information exist already . is it a job of organizing parsing of existing information (MS already stores some info)

11. Italian Radio Data Archive Alessandra Zanichelli
08/09/2020, 11:50

VLBI-it with 3 telescopes
* entities:

    * single dish data (FITS), pulsar single dish/VLBI (PSRFITS), VLBI (UVFITS) - rk: FITS-IDI is generated
    * exposure + metadata
    * schedule and logfiles

* derivations between products kept, would need to describe processing pipeline steps (calibration in particular)
* some provenance info in headers

    * context
    * quality

Features for each project : (copy into each description if relevant and complete)
IVOA Provenance concepts reused in the use-case

- entities
- agents
- activities
- parameters
- used/wgby relations
- descriptions of entities
- description of activities
- dependency relations : was derived from , was informedby

Provenance peculiar aspects:
capture/recording steps
granularity : collections of datasets / coarse/fine representation for tasks(activities)
activity descriptions stemming from Workflow definition ?
provenance graph: generated on the fly from serialization ? stored in DB? graph interface?

Discussion topics:

* how to **serialize** provenance:
    * demo of the (vo)prov Python package (W3C files)
    * YAML serialization explored for Vizier catalogs
* mapping of the information available to this serialization ("on-top" provenance)
* capture provenance "inside", example of ctapipe (a pipeline framework)
* create a channel on the ESCAPE chat (or slack?), gitlab page
* how to move from provenance "on-top" to provenance "inside"?
* relation between workflow description (CWL) and provenance records?

[MF]
some questions that CTA is concerned with at the moment and that are linked to this workshop:

- **minimum** amount of information in provenance +5 subject[ïï]

    - for end user (science user): users of observatory data, make physics results reproducible
    - for observatory staff: responsible for complex data processing and calibration chain, to deliver high-quality data to end users, for data archives spanning decades of data
    - needs towards provenance information is different in the two cases
- level of **granularity** for provenance information
    - at the level of observation? at the level of GTI? at the level of file?
- where to **store** provenance information? +4
    - as part of the data itself
    - in a dedicated repository / metadata catalogue / DB
    - in both
    - additional topics touch questions of synchronisation, updating/removing information etc
- where to **track** provenance information?
- how to **link** provenance information
    - not all provenance information can become part of the file header (too long, too much information, uses internal information)
    - how to best link provenance information as part of the data product to some additionally available provenance information (e.g. usage of UIDs?)
    - example: instrument configurations linked from provenance information
    - where to **host** / **attach** the provenance metadata DB
    - is a provenance metadata function of the archive or of the data processing workflow machinery or of the science portal (with or without jupyter notebook hub)
- standardised descriptions of complex data processing **workflows** and provenance (A)
    - best practices and standards should be possible to map data processing workflows, so can't the provenance data model extended to contain standards for workflow descriptions?
    - data processing workflows can be very complex, not always a chain of steps, can include merging of outputs as input for the next steps etc., how to map these in provenance information? can we create examples / templates for such workflows mapped into provenance?
- provenance in **VOEvent**
    - what is the minimum provenance information needed in VOEvent to enhance multi-messenger follow-up observations and establish the concepts of FAIR in the multi-messenger context?

- FB: How to give acces to the provenance of a single measurement or row in a measurement table (or event)
- FB : storing in nonSQL databases may introduce interoperability problem. TAP doesn't work on top of these databases.

ML: can we bind together provenance records created from different data centers merging steps executed across a distributed workflow at various data centers? +1

KM3NeT [JS]:

- Interested in provenance for complex workflows and Grid computing, especially involving DIRAC (overlap with CTA)
    - provenance **database** versus **infile** provenance +1 --subject **[I]** ?
    - configuration information and annotation in complex workflows and use in mass processing (use of Common Workflow Language?)+1
    - How to do provenance when workflow is defined "on the fly" (data driven workflows) [goes together with 'A' in Matthias list]+1
- (aside) Reference on conceptualization of "provenance": https://www.youtube.com/watch?v=wyt0Zhbd1T0 (BD2K presentation on Provenance)

Notes on the discussion (afternoon session)
provenance database versus infile provenance
HV: infile is more than nothing. may be the first step?
pb of interoperability if the prov language is not

JS: What is the interesting part of provenance , (and for which science, ML)
what is the part of data management
the line between this 2 categories is tiny

project 's internal need / user's need

Marco : minimum amount of information that Provenance should provide?
needed
interoperable

MF: different types of users: data providers (data managers), end users (scientists, astronomers) , etc. because they have different needs
can we define the diff. types? can be defined in the specification ?
proposed terms :
internal provenance --> more detailed, DB?
user's provenance (to know what the data are and how they can be used) --> less detailed, minimum? infile? .
information in the header of the FITS file?

config and parameters
clarify the language :
instruments settings , processing parameters,

FB: in terms of data discovery
attached/detached to dataset information depends on how you discover data sets : either you discover the dataset first and then find out attached provenance information or we discover datasets in database by constraining provenance information

MK: can decide the minimum depending of what's the user wants to view on a graph
filter on the levels : deeper level , intermediate, coarse

CB (and ML) : draw the line between science requirements and data managenent in the past the distinction between minimal and full details was delivered by the format
FITS header, in ascii , human readable format was a summary of important science features.

TG: should we go for a specification for a minimal profile
with defined keywords, create a validator for serialization documents whether they comply to the minimal profile those words,
JS: from the discoverability , we can try to standardize the vocabulary (provCore?)

Marco: Core table as a view in PROV tap : we could elaborate from uses-cases , and sets of queries we consider all users may be interested in (and services should provide)

GL: for Vizier a simple table for a coarse prov metadata profile would be appreciated
vocabulary ? other description formats ?

JS: what is the follow-up of this forum in terms of specifying this minimal Provenance profile?
MS: which communication channel to propose
JS: using git Instance git lab at in2P3
MF: project : list of items common to many projects we can discuss
, but we could also organise dedicated workshops on topics : tech forums ? get the expert people involved, produce a practical provenance implementation

Deliverables could be : specification of minimal PROV profile (e.g. as IVOA note?), PROV tools library, serialization validator, etc.
MA: for visibility : provide feedback at IVOA meetings, ADASS posters, publications in 'Astronomy and Computing'
for uptake : concrete tool boxes elaborated in tech forums (focused workshops)
intersections with wP3 and WP5 in the Escape project

MS: wants to focus on practical results to be able to produce, circulate and handle provenance serialization

connection to IVOA:

MS: focus to follow? what de we want to show to the user?
when data levels are well defined , and granularity of activities too, we should be able to decide about this common profile

How can we store Provenance representation
example LST (JRE): a structured db which organises PROV information in folders
any experience with noSQL DB? graph databases ?
DB architecture not yet considered for some projects

KM3net : very much connected to distributed workflows
MF: CTA also looking for such complexity : multiple data centers, complex WF
how are systems intertwined across datacenter , user's types, etc...

FB: does anybody experienced translation from graphs to tabular DB ? (for example using CTE - Common table expressions-)

Marco: CWL used in GAPS
CTA DIRAC / Michèle S 's prototype: PROV is stored in its own "external" DB (external but on a CTADIRAC server)

TamàsG: used in KM3net
have experienced a representation in CWL
docker , HPC , compatible

JS: https://git.km3net.de/tgal/pipelines -> tools, -> examples

Mattia Mancini : In LOFAR in the EOSC project we started converting some standard pipeline for calibration and preprocessing. We also started to create CWL wrapper for standard tool used to process LOFAR data
The reference repo are at https://git.astron.nl/eosc/lofar-cwl/ for the steps andhttps://git.astron.nl/eosc/prefactor3-cwl

MK : in CWL you cannot make decision in the execution of the WF.
sequential for the moment .
=> sceptical about CWL

TG : used to description of WF . but not really able/useful to optimise the execution
MattiaM: but efficient for description of the steps and what they consume.

MF: CTA is exploring the use of CWL for the pipelines

MF: running many telescopes simultaneously : workflow to maintain on operational mode.

MServillat: examples of a document in a dedicated Provenance YAML format / SVG graph translation
conversion is easy/ format is easy to read
Translation possible to other formats : not yet part of the voprov library, but could be part of it .

MS : how to go further ?
list of topics / tasks to explore further

FB proposal for a short synthesis of the discussion :
We can sort out provenance needs into two categories : provenance for managment and provenance for science. The level of graining of provenance information depends from this categorization.

Beside this come a couple of questions:

One of the question behind this is : do we integrate all workflow details in Provenance and how do we do it (mapping or integrating CWL in provenance? others workflow languages ?)

Another major question is do we join provenance information to the datasets (eg in headers) or do we store it in databases. Do we do both ?

What is the main access flow direction between provenance information and datasets : do we want provenace information associated to a datasets or do we want to select datasets according to their provenance features ?

Main structure of provenance is graph oriebted. How can we user-friendly query provenance databases and visualize the results ?

Full provenance is complex : what kind of simplified views can we provide and for which usage

Personal conclusion : where do we need access protocols refinemnt? where do we need new tools (applications) ? where do we simply need implementation description, help, faq ?